DOI Link: 10.14421/impulse.2025.52-05

IMPULSE: Journal of Research and Innovation in Physics Education
Volume 5, Issue 2, 138 – 151

© ISSN (p): 2798–1762; ISSN (e): 2798-1754
http://ejournal.uin-suka.ac.id/tarbiyah/impulse

Quality Analysis of the Final Examination Instrument and Its Implication for Student Learning Outcomes in the Fundamentals of Physics II

Prawitasari*, Purbo Suwasono, Sutopo, Ratika Sekar Ajeng Ananingtyas, Dyah Palupi Rohmiati

Departement of Physics, Universitas Negeri Malang, Malang, Indonesia
*Corresponding author: prawitasari.fmipa@um.ac.id

ABSTRACT

The low student learning outcomes in the Fundamentals of Physics II raise concerns about the quality of the evaluation instrument used to measure the students' achievement. This study aims to analyze the quality of the Final Semester Examination (UAS) instrument and its relation with students' learning outcomes. The subjects of this study were 53 second-semester students from the Department of Physics, Universitas Negeri Malang. The evaluation instruments consisted of 30 multiple-choice questions. The items were analyzed using PSPP statistical software, which functions similarly to SPSS. The analysis involved validity testing using the Pearson Correlation Method, reliability testing using Cronbach's Alpha, the discrimination power and difficulty indices. The results showed that 50% of the items were valid, the instrument's reliability was moderate with a Cronbach's Alpha value of 0.60, and only 10% of the items had very good discrimination power, while most (63.4%) were categorized as low to negative. In terms of difficulty, 53.3% of the items were classified as difficult, 20% as moderate, and 26.7% as easy. These findings indicate that the low student learning outcomes are influenced not only by internal factors but also by the quality of the evaluation instrument. So, this study not only test the validity and reliability of the test items but also explores how the question construction can affect learning achievements. This study offers new insights for lecturers to design more accurate instruments that can truly reflect students' ability, especially in the Fundamentals of Physics II course.

INTISARI

Rendahnya capaian hasil belajar mahasiswa pada matakuliah Fisika Dasar II menimbulkan kekhawatiran terhadap kualitas intrumen evaluasi yang digunakan untuk menilai hasil belajar. Penelitian ini bertujuan menganalisis kualitas butir soal Ujian Akhir Semester (UAS) dan hubungannya dengan hasil belajar mahasiswa. Subjek dari penelitian ini adalah 53 orang mahasiswa semester kedua pada Departemen Fisika, Universitas Negeri Malang. Instrumen yang dievaluasi adalah soal UAS yang berjumlah 30 soal berbentuk pilihan ganda. Soal di analisis dengan perangkat lunak PSPP. PSPP merupakan perangkat

ARTICLE HISTORY
Received: October 14, 2025
Accepted: November 18, 2025

KEYWORDS:
Difficulty Indices, Discriminatory Power, Fundamentals of Physics, Reliability, Validity

KATA KUNCI:
Daya Beda, Fisika Dasar, Kesukaran Soal, Reliabilitas, Validitas

* Corresponding author:
Prawitasari, Department of Physics, Universitas Negeri Malang, Indonesia
prawitasari.fmipa@um.ac.id +62 85-6664-0074

lunak statistik yang memiliki fungsi mirip dengan SPSS. Soal dianalisis dengan Uji Validitas menggunakan Pearson Correlation Method, Uji Reliabilitas dengan Cronbach's Alpha, daya beda soal dan tingkat kesukarannya. Hasil analisis menunjukkan 50% soal valid, reliabilitas instrumen moderat dengan Cronbach's Alpha 0,60, serta hanya 10% soal memiliki daya beda sangat baik, sedangkan sebagian besar (63,4%) tergolong rendah hingga negatif. Dari sisi kesukaran, 53,3% soal sukar, 20% sedang, dan 26,7% mudah. Temuan ini mengindikasikan bahwa rendahnya hasil belajar mahasiswa tidak hanya dipengaruhi faktor internal, tetapi juga kualitas instrumen. Sehingga, penelitian ini tidak hanya menilai validitas dan reliabilitas soal, tetapi juga menelusuri bagaimana konstruksi soal dapat mempengaruhi capaian belajar. Penelitian ini juga memberikan sudut pandang baru bagi dosen untuk merancang instrumen yang dapat merefleksikan kemampuan mahasiswa, khususnya pada matakuliah Fisika Dasar II.

A. Introductions

Fundamentals of Physics is one of the course that must be completed by first semester students in the Department of Physics, Universitas Negeri Malang, both in education and non education program. This course is really important because it's being a foundation for all advanced physics subjects and shaping students' conceptual understanding. Students are expected to connect the basic theory with mathematical formulations and know how to applying it in real life phenomena[1],[2]. Especially for students in education programs, well understanding the basic concept in Fundamentals of Physics also prepares them to become a teacher who can explain physics phenomena to others effectively. Because of this, students' achievement in this course reflects how successful the learning process.

From the evaluation of learning outcomes of Physics Education students in Department of Physics, Universitas Negeri Malang, the results of the Final Semester Examination (UAS) in the Fundamentals of Physics II course were found relatively low. This condition raises questions about the problems in the learning process, such as the effectiveness of teaching methods, students' understanding, and also the instrument's quality that are used to evaluate students' learning outcomes [3],[4],[5]This case indicates a gap between the learning objectives stated in the Course Learning Outcomes (CPMK) and the results students actually achieve. Low examination results do not always mean that students lack of understanding. But, they also can reveal weaknesses of the test items. When test instuments are not properly validated, it may be leading to inaccurate interpretation of students' abilities.

The assessment process cannot be separated from the quality of the assessment instruments used [6],[7]. A good instruments, through written or oral tests, assignments and project based, should be able to measure cognitive, psychomotor and affective skills[3],[8]. In physics learning, test items play a role in identifying how well students in understanding the basic concepts and how to applying it in various problem contexts [9],[10],[11],[12]. Because of that, the quality of test items that used in evaluation need to be analyzed to ensure that the given accurately, objectively, and proportionally measure students' abilities[13],[14],[15],[16].

A good assessment instrument should have a balanced level of difficulty and non-ambiguous answer choices[17]. An effective instrument must be able to measure the achievement of the competencies stated in the Learning Outcomes. Problematic instruments, such as those that are too difficult or contain ambiguous wording, can cause evaluation outcomes that do not accurately reflect students' actual abilities. To determine the causes of low student learning outcomes, it is necessary to analyze the test items used in the Final Semester Examination (UAS). Items analysis helps to identify the quality of each question, whether it is good, less effective or poor, or which questions perform well and which need revision, so it can be used as the assesment instruments[18]. A test item is considered high quality if it has strong validity, good reliability, and meets other criteria outlined in the assessment guidelines [3],[15],[19].

Moreover, item's analysis can give information how each question distinguishes between students with high and low levels of understanding effectively. This process, helps lecturer to identify items that need to be revised, removed or replaced. Item's analysis also can identify the strengths and weaknesses of each question, give detailed information about test items, and identify the problem within the questions themselves[3],[20],[21].

An important aspect that being urgency in this study is the test items used in the Final Semester Examination has not passed validity and reliability testing before being distributed to students. So, the test instrument does not guarantee that it can measure students' abilities objectively. Based on that need, this study focuses on analyzing the multiple choice test items used in The Fundamental of Physics II Final Semester Examination to evaluate students' learning outcomes. The analysis of test items cover validity and reliability testing, discriminatory power and difficulty indices. This study expected to provide useful insight for improving the quality of assessment instrument. Also, it can serve as a reference for the lecturers to make assessment process is fair, objective and students' learning outcomes truly reflective.

B. Methods

General Background

This study uses descriptive quantitative methods to analyze the evaluation instrument in the Final Semester Examination (UAS) items in the Fundamentals of Physics II and its relation to students' learning outcomes. The focus of this study was to assess the validity and reliability of the test items, the discriminatory power dan the difficulty indices. This approach was chosen because the 4 aspects provides empirical evidence, ensuring that the evaluation instrument able to measure the learning outcomes consistently, objectively and proportionally.

Participants

The participants of this study were 53 second-semester students enrolled in the Fundamentals of Physics II course during the even semester of the 2024/2025 academic year, in the Department of Physics, Universitas Negeri Malang.

Instruments and Procedures

The instrument used in this study was Final Semester Examination sheet, containing 30 multiple-choice questions designed to assess the level of students' understanding of the fundamental concepts. Each question was scored : a score of 1 if the answer is correct and a score of 0 if the answer is incorrect. The students' responses were recorded on standard answer sheets.

Each question was constructed to measure one or more indicators in the Sub-Course Learning Outcome (Sub-CPMK). For Example :

Sample Question 1 (Sub CPMK 3.3 : Students showed mastery the Thermodynamics concepts and able to identify and analyze solutions for standard physical systems.)

A parfume bottle made from glass, consists of a main bottle and a vertical cube as shown in the Figure 1.

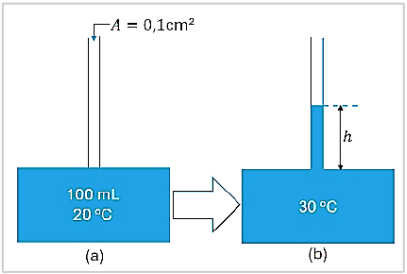

Figure 1. Liquid parfume fulfills the main bottle (a) and parfume rise into the tube by a height of *h* (b)

In room temperature (20 °C), 100 cm^3 liquid parfume poured into the bottle, it fulfills the main bottle (Figure 1.a). The vertical tube has cross-sectional area 0.1 cm^2. When the bottle heated to a temperature of 30 °C, parfume from the main bottle will rise into the tube by a height of *h* (Figure 1.b). Assume that the linear expansion coefficient of the glass is $0.3 \times 10^{-6}/℃$ and the volume expansion coefficient of the liquid parfume is $1.5 \times 10^{-4}/℃$. What is height *h* ?

a. 1.5 mm
b. 5 mm
c. 10 mm
d. 15 mm
e. 16.5 mm

Sample Question 2 (Sub CPMK 3.4 : Students showed mastery the Optics concepts and able to identify and analyze solutions for standard physical systems.)

Two symmetric biconcave lenses A and B, made from the same material but have different radius of curtivature and also have different thicknesses $t_B > t_A$ (see Figure 2 below)

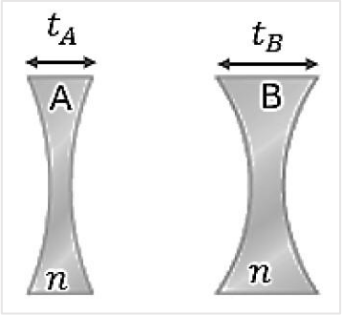

Figure 2. Two symmetric biconcave lenses A and B

The ratio of the radius of curtivature *R* and the focal lengths *f* of two lenses are…

a. $R_B > R_A$ and $f_B > f_A$
b. $R_B > R_A$ and $f_B < f_A$
c. $R_B < R_A$ and $f_B < f_A$
d. $R_B < R_A$ and $f_B = f_A$
e. $R_B < R_A$ and $f_B > f_A$

Data Analysis

The data were analyzed using PSPP, a statistical software package that functions similarly to SPSS. Using PSPP, validity testing (Pearson correlation), reliability testing (Cronbach's Alpha), discriminatory power, difficulty indices were computed for each question. The test item was analyzed in four stages :

1. Validity test basically means "measure what is intended to be measured"[22]. The validity testing uses the Bivariate Pearson (Product Moment) correlation at a 5% significance level to determine the accuracy of each test item. With 53 respondents, the r-table used as the threshold was 0.266. Items with score equal to or greater than the threshold were categorized as valid and it contributed to assessing students' learning outcomes. The validity value was computed using Bivariate Pearson formula, expressed as [13], [20], [23]:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}} \quad (1)$$

2. Reliability tes refers to the consistency of a research study or measuring test[24]. Reliability testing uses the Cronbach's Alpha coefficient, with interpretation: 0.00 – 0.20 (very low), 0.20 – 0.40 (low), 0.40 – 0.60 (medium), 0.60 – 0.80 (high) , and 0.80 – 1.00 (very high)[3],[25] ,[26] ,[27]. The instrument considered as reliable if the Cronbach's Alpha value is greater than 0.60[28]. In this analysis,

the reliability value was computed using the Cronbach's Alpha formula, expressed as [23],[28],[29]:

$$r=\left(\frac{n}{n-1}\right)\left(\frac{S^2-\sum pq}{S^2}\right) \quad (2)$$

$$S^2=\frac{\sum x^2-\frac{(\sum x)^2}{n}}{n} \quad (3)$$

3. Discriminatory power analysis to evaluate the ability of each question to distinguish between high and low-achieving students, with criteria ranging from 0.70–1.00 (excellent) to negative values (poor)[3],[23]. The discriminatory power value was computed using formula below [13], [23]:

$$D=P_A-P_B=\frac{B_A}{J_A}-\frac{B_B}{J_B} \quad (4)$$

4. Item difficulty analysis or difficulty indices, which classifies test items according to their level of difficulty based on the proportion of students who answered each question correctly[23],[30]. According to standard interpretation criteria, items are categorized as difficult (0.00–0.30), moderate (0.31–0.70), or easy (0.71–1.00)[31]. The formula used to analyze the item difficulty is [13],[23]

$$P=\frac{B}{J_s} \quad (5)$$

C. Result and Discussion

Result

Validity Test

In analyzing the data using PSPP, the first stage was testing the validity of the Final Semester Examination (UAS) items. The results of the validity testing are presented below on Table 1.

Table 1. Validity Test Result

Items	Result	Criteria	Items	Result	Criteria	Items	Result	Criteria
1	0.309	V	11	0.014	IV	21	0.276	V
2	0.301	V	12	0.377	V	22	0.394	V
3	0.127	IV	13	0.015	IV	23	0.121	IV
4	0.289	IV	14	0.258	IV	24	0.499	V
5	0.059	IV	15	0.256	IV	25	0.049	IV
6	0.164	IV	16	0.526	V	26	0.208	IV
7	0.277	V	17	0.535	V	27	0.217	IV
8	0.403	IV	18	0.692	V	28	0.398	V
9	-	IV	19	0.445	V	29	0.511	V
10	-	IV	20	0.519	V	30	0.403	V

Note: V = Valid, IV = Invalid

Reliability Test

The second stage was conducting the reliability test. The reliability test used to measure the consistency of an assessment instrument. Based on the analysis using the PSPP program with the Cronbach's Alpha coefficient, the results presented below on Table 2.

Table 2. Reliability Test Result

Case Processing Summary			Reliability Statistics	
Cases	**N**	**Percent**	**Cronbach's Alpha**	**N of Items**
Valid	53	100.0 %	0.60	30
Excluded	N	.0%		
Total	54	100.0%		

Discriminatory Power

The third stage is measuring the discriminatory power of each question. In this study, the discriminatory power was measured using the *Corrected Item-Total Correlation* value in the PSPP program. This correlation value represents the strength of the relationship between each item's score and the students' total score. The result of discriminatory power each showed below on Table 3.

Table 3. Discriminatory Power Test Results

Item-Total Statistics			
Items Number -	**Scale Mean if Item Deleted-**	**Scale Variance if Item Deleted-**	**Corrected Item-Total Correlation**
1	11.57	10.29	0.19
2	11.72	10.51	0.23
3	11.66	10.81	0.02
4	11.62	10.39	0.18
5	11.64	10.97	-0.05
6	11.04	10.69	0.03
7	11.34	10.31	0.13
8	11.36	9.89	0.27
9	11.79	10.98	NaN
10	11.79	10.98	NaN
11	11.36	11.27	-0.16
12	11.53	10.06	0.25
13	11.74	11.01	-0.05
14	11.62	10.47	0.15
15	11.70	10.56	0.17
16	10.87	10.12	0.46
17	10.94	9.82	0.45
18	11.06	9.13	0.61
19	11.04	9.88	0.33

Item-Total Statistics			
Items Number -	Scale Mean if Item Deleted-	Scale Variance if Item Deleted-	Corrected Item-Total Correlation
20	10.94	9.86	0.43
21	11.36	10.31	0.13
22	11.08	9.99	0.27
23	11.49	11.56	-0.25
24	11.47	9.64	0.38
25	11.15	11.05	-0.10
26	11.34	10.54	0.06
27	11.70	10.64	0.13
28	11.66	10.19	0.31
29	10.89	10.06	0.44
30	11.53	9.98	0.28

Item Difficulty Result

The last stage in analyzing the test instrument is the level of item difficulty or difficulty indices. The difficulty indices calculated based on the average number of student who answered each question correctly or mean. The result of difficulty indices analysis is showed on Table 4 below.

Table 4. Difficulty Indices Test Result

Items Number -	N		Mean	Standard Deviation	Minimum	Maximum
	Valid	Missing				
1	53	0	0.23	0.42	0.00	1.00
2	53	0	0.08	0.27	0.00	1.00
3	53	0	0.13	0.34	0.00	1.00
4	53	0	0.17	0.38	0.00	1.00
5	53	0	0.15	0.36	0.00	1.00
6	53	0	0.75	0.43	0.00	1.00
7	53	0	0.45	0.50	0.00	1.00
8	53	0	0.43	0.50	0.00	1.00
9	53	0	0	0	0.00	1.00
10	53	0	0	0	0.00	1.00
11	53	0	0.43	0.50	0.00	1.00
12	53	0	0.26	0.45	0.00	1.00
13	53	0	0.06	0.23	0.00	1.00
14	53	0	0.17	0.38	0.00	1.00
15	53	0	0.09	0.30	0.00	1.00
16	53	0	0.92	0.27	0.00	1.00
17	53	0	0.85	0.36	0.00	1.00
18	53	0	0.74	0.45	0.00	1.00
19	53	0	0.75	0.43	0.00	1.00

Items Number -	N		Mean	Standard Deviation	Minimum	Maximum
	Valid	Missing				
20	53	0	0.85	0.36	0.00	1.00
21	53	0	0.43	0.50	0.00	1.00
22	53	0	0.72	0.45	0.00	1.00
23	53	0	0.30	0.46	0.00	1.00
24	53	0	0.32	0.47	0.00	1.00
25	53	0	0.64	0.48	0.00	1.00
26	53	0	0.45	0.50	0.00	1.00
27	53	0	0.09	0.30	0.00	1.00
28	53	0	0.13	0.34	0.00	1.00
29	53	0	0.91	0.30	0.00	1.00
30	53	0	0.26	0.45	0.00	1.00

Discussion

Validity Test

The validity test for the Final Semester Examination in the Fundamentals of Physics II course ensures that the test items can measured students' abilities accurately in accordance with the course learning outcomes. With a total of 53 respondents, the validity testing used the Pearson Product-Moment correlation between each item score and the total score, applying a 5% significance level. In this condition, 0,266 used as the r-table value.

Based on the 30 test items analyzed on Table 1, 15 (50%) were declared valid: 1, 2, 7, 12, 16, 17, 18, 19, 20, 21, 22, 24, 28, 29, and 30. Meanwhile, the other 15 items (50%) were found to be invalid, including items numbers 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 15, 23, 25, 26, and 27. From these results, we can see that only half of the questions met the validity criteria, and the remaining items couldn't be used to measure students' learning outcomes accurately.

The valid test showed high validity can be found in item 17 which had a correlation coefficient of r = 0.535 and item 18 with r = 0.692. The item score with values about 0.50 indicate high validity. It's mean that these items can be used as assessment instrument in future and also can effectively distinguish between students who have understanding the course and those who have not. In the other side, items called invalid when they have very low or negative correlation value. For example, item 13 (r = 0.015) and item 11 (r = -0.014). These items are contributing little to measuring students' ability. The low validity of test items can be caused by several factors such as ambiguous answer choices, unclear question wording, and students' misconception about the materials.

A more extreme condition was observed in items 9 and 10, which could not be analyzed for validity because none of the students answered them correctly (all respondents scored 0). With no variation in the item score, it becomes statistically

imposible to compute the correlation coefficient. It called as a dead item in Classical Test Theory. The questions in this type need to be revised immediately because they have no meaningful information about students' ability.

Overall, we can concluded that the invalid test items need to be reviewed to ensure that the assessment function as intended. The development of valid assessment instrument is important to ensure that evaluation results accurately reflect the abilities that being measured.

Reliability Test

Overall, the reliability analysis in Table 2 above, yield a value of 0.60. This value is below the ideal threshold of 0.70, which is generally used for a reliable instrument. However, it is still in the moderate or acceptable reliability category and can be used to assess learning outcomes[27] .Although the Cronbach's Alpha value is 0.60, the test's quality still needs improvement. A qualitative review of the items, refinement of question wording, and the selection of more effective distractors can help enhance the overall reliability and quality of the instrument. This is also because, even if we say an instrument may be reliable, it is not necessarily valid, because reliability must be combined with validity[4].

Discriminatory Power

Discriminatory power is one of the main indicator of test quality, as it shows an item's ability to distinguish between students with high and low levels of understanding. The higher the discriminatory power value, the more effective the item in identifying students who have truly mastered the material. In general, the interpretation of correlation values follows the guideline that ≥ 0.40 are considered as very good, 0.30–0.39 are good, 0.20–0.29 are moderate and 0.00–0.19 are poor. Negative values are regarded as very poor and unsuitable for use. The results of discriminatory power tes can we see on Table 3. Based on the analysis of 30 questions from the Fundamentals of Physics II Final Examination with a total of 53 respondents, only 3 items (10%) were categorized as very good, specifically items 20, 27, and 29. This indicates that only a small portion of the questions effectively distinguishes students with high mastery of the material. Meanwhile, 2 items (6.7%) were classified as good (items 28 and 30), and 4 items (13.3%) were classified as moderate.

However, a rather concerning finding is that the majority of the items (63.4%) fall into low to very low discriminatory power category. Fourteen items were identified as having low discriminatory power, while five showed negative correlation values. Items with low or negative discriminatory power undermine the test overall validity and may lead to misleading interpretations of the results. When an item produces a negative discrimination value, it's indicating that the items not functioning as intended. Items with negative discriminatory power values should be revised immediately, because they can undermine the test's overall reliability.

Item Difficulty Result

Based on the table 4, difficulty indices test results, most questions were categorized as difficult. 16 items (about 53.3%) had difficulty indices of < 0,30, including 1, 2, 3, 4, 5, 9, 10, 11, 12, 14, 15, 21, 26, 28, 29, and 30. This finding indicates that less than 30% of students were answered correctly and also these 16 items were found relatively challenging for students and may need to be revised in terms structure of wording, cognitive level and content alignment.

A total of 6 items (about 20%) were categorized as moderate, including 7, 8, 22, 23, 24 and 25. It's mean that they have difficulty indices ranging from 0,31 to 0.70. Every items in this category are classified as ideal items because the can distinguish proportionally between students with high and low understanding. Last, 8 items (about 26.7%) were categorized as easy, including 6, 16, 17, 18, 19, 20, 22, and 27. These items have difficulty indices about 0.70. The amount of easy items is still acceptable, if too many, it can reduce the test's overall discriminatory power.

The unbalanced composition of difficulty levels indicates that improvements to the question bank are necessary, especially for items with difficulty levels. Questions that are too difficult can cause frustration among students and reduce the validity of the evaluation results, while questions that are too easy may obscure students' true abilities. Therefore, a thorough evaluation and revision of items with excessively high or low difficulty indices is strongly recommended to improve the test instrument's quality for future exams.

Overall, this study shows that the low achievement in student learning outcomes is not only caused by students' internal factors, but is also related to the quality of the evaluation instruments used. To improve students' learning outcomes, actions incude enhancing teaching methods, providing relevant practice questions, and developing of evaluation instruments that are valid, reliable, have high discriminatory power, and have a balanced distribution of difficulty levels.

D. Conclusion

Based on the study's result, the evaluation instrument used in the Fundamental of Physics II Final Examination still needs improvement. Only half of the questions were found to be valid; the instrument's reliability was in the moderate range, and most of questions showed low discriminatory power and an unbalanced level of difficulty proportion. This condition indicates that the evaluation instrument does not fully effective in representing students' ability accurately. The low student's learning outcomes observed in this course does not fully reflect their actual abilities, but also influenced by the weaknesses of the evaluation instrument itself. When the test items too difficult or have ambiguous phrased, the test result cannot be used as an accurate representation of students' understanding of the material. The findings underline that the evaluation instrument need to be urgently improved. In the future, assessment instrument should focus on increasing the validity and realibity, balancing the

difficulty level, and enhancing discriminatory power. Developing better instrument will help to produce more accurate and more informative evaluations, also improving the quality of students' learning outcomes.

Acknowledgements

The author would like to express sincere gratitude to the Department of Physics, Universitas Negeri Malang, for the support and facilities provided during the course of this research. Appreciation is also extended to the participating students and fellow lecturers for their support in completing this study.

References

[1] Hamatun and M. R. Rifai, "Studi Pemahaman Konsep Energi dalam Penyelesaian Berbagai Persoalan Fisika Pada Perkuliahan Fisika Dasar," *Al-Ikmal: Jurnal Pendidikan*, vol. 2, no. 1, pp. 90–99, 2022.

[2] D. Riwanto, A. Azis, and K. Arafah, "Analisis Pemahaman Konsep Peserta Didik dalam Menyelesaikan Soal-Soal Fisika Kelas X MIA SMA Negeri 3 Soppeng," *JSPF*, vol. 15, no. 2, Nov. 2019, doi: 10.35580/jspf.v15i2.11033.

[3] H. D. Saputra, W. Purwanto, D. Setiawan, D. Fernandez, and R. Putra, "Hasil Belajar Mahasiswa: Analisis Butir Soal Tes," *Edukasi: Jurnal Pendidikan*, vol. 20, no. 1, pp. 15–27, 2022, doi: 10.31571/edukasi.v20i1.3432.

[4] A. Loka Son, "Instrumentasi Kemampuan Pemecahan Masalah Matematis: Analisis Reliabilitas, Validitas, Tingkat Kesukaran Dan Daya Beda Butir Soal," *Gema Wiralodra*, vol. 10, no. 1, pp. 41–52, 2019, doi: 10.31943/gemawiralodra.v10i1.8.

[5] A. Iskandar and M. Rizal, "Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 22, no. 1, pp. 12–23, June 2018, doi: 10.21831/pep.v22i1.15609.

[6] Y. Utami, "Uji Validitas dan Uji Reliabilitas Instrument Penilaian Kinerja Dosen," *SAINTEK*, vol. 4, no. 2, pp. 21–24, Feb. 2023, doi: 10.55338/saintek.v4i2.730.

[7] R. Ika Ningtiyas, M. Sahal, and Y. Gusmeri, "Validitas dan Reliabilitas Butir Soal Berbasis Kemampuan Berpikir Kritis Pada Materi Listrik Dinamis," *j.armada.pendidik.*, vol. 1, no. 1, pp. 26–30, Jan. 2023, doi: 10.60041/jap.v1i1.5.

[8] K. Bashooir and S. Supahar, "Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran fisika berbasis STEM," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 22, no. 2, pp. 219–230, Dec. 2018, doi: 10.21831/pep.v22i2.19590.

[9] C. H. Sulistiawan, "Kualitas Soal Ujian Sekolah Matematika Program IPA dan Kontribusinya Terhadap Hasil Ujian Nasional," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 20, no. 1, pp. 1–10, June 2016, doi: 10.21831/pep.v20i1.7516.

[10] A. Anita, S. Tyowati, and Z. Zuldafrial, "Analisis Kualitas Butir Soal Fisika Kelas X Sekolah Menengah Atas," *J Educ*, vol. 16, no. 1, p. 35, June 2018, doi: 10.31571/edukasi.v16i1.780.

[11] Nasir, “Analisis Empirik Program Analisis Butir Soal Dalam Rangka Menghasilkan Soal Yang Baik dan Bermutu Sebagai Alat Evaluasi Pembelajaran Fisika,” Pros. Semirata, 2015, pp. 336–347.
[12] Umacina, “Analisis butir soal sumatif semester ganjil mata pelajaran fisika,” *J. Pendidik. Fis. Unima*, vol. 1, no. 2, pp. 33–38, 2020.
[13] F. R. Dianova and N. Anwar, “Analisis Butir Uji Validitas, Reliabilitas, Tingkat Kesukaran, dan Daya Pembeda Soal Sumatif Bahasa Arab SD Islam,” *Jurnal Bahasa Daerah Indonesia*, vol. 1, no. 3, p. 13, 2024, doi: 10.47134/jbdi.v1i3.2863.
[14] M. Solichin, “Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan,” *Dirāsāt J. Manaj. Pendidik. Islam*, vol. 2, no. 2, pp. 192–213, 2017.
[15] T. Novia, A. Wardani, C. Canda, N. Nurdi, and N. Nurmasyitah, “Analisis Validitas dan Reliabilitas Butir Soal UTS Fisika Kelas X SMA Swasta Muhammadiyah 4 Langsa,” *GRAVITASI: Jurnal Pendidikan Fisika dan Sains*, vol. 3, no. 01, pp. 19–22, 2020, doi: 10.33059/gravitasi.jpfs.v3i01.2256.
[16] L. Purniasari, M. Masykuri, and S. R. D. Ariani, “Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia SMA N 1 Kutowinangun Tahun Pelajaran 2019/2020 menggunakan Model Iteman dan Rasch,” *J. Pendidik. Kim.*, vol. 10, no. 2, pp. 205–214, 2021.
[17] N. Abdullah, M. Jahja, and D. G. E. Setiawan, “Analisis Kualitas Butir Soal pada Mata Pelajaran Fisika di Jurusan Fisika Fakultas MIPA Universitas Negeri Gorontalo Tahun Ajaran 2021/2022,” *JSPF*, vol. 18, no. 1, p. 44, Apr. 2022, doi: 10.35580/jspf.v18i1.32358.
[18] Fitrah, “Analisis Butir Soal Ulangan Akhir Semester Ganjil Mata Pelajaran Teori Kejuruan Akuntansi Analysis of the Final Examination Items of Teori Kejuruan Akuntansi At Fisrt Semester,” pp. 1–11, 2016.
[19] “Kualitas Soal Bahasa Indonesia di SMP Muhammadiyah 1 Pontianak: Analisis Butir Soal,” *J. Pendidik. Bhs. dan Sastra Indones.*, vol. 11, no. 2, pp. 112–119, 1022.
[20] I. Magdalena, A. Fitroh, D. K. Fadhilah, D. Habsah, and R. Y. Qodrawati, “Mengelolah Data Uji Validitas Ddn Reliabilitas Dalam PenelitianPendidikan : Instrumen Tes Dan Non Tes Peserta Didik Kelas IV SDN Pondok Kacang Barat 03,” *Jurnal Pendidikan Sosial Dan Konseling*, vol. 01, no. 02, pp. 49–53, 2023.
[21] Y. F. Alista and R. A. Syahzanani, “Analisis Butir Soal Ulangan Harian Fisika dengan Pendekatan Teori Tes Klasik menggunakan Program Anates,” pp. 1–11, 2023.
[22] H. Taherdoost, “Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research,” *Jurnal Elektronik SSRN*, 2016, doi: 10.2139/ssrn.3205040.
[23] I. Magdalena, S. N. Fauziah, S. N. Faziah, and F. S. Nupus, “Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas Iii Sdn Karet 1 Sepatan,” *BINTANG : Jurnal Pendidikan dan Sains*, vol. 3, no. 2, pp. 198–214, 2021.
[24] I. Kennedy, “Sample Size Determination in Test-Retest and Cronbach Alpha Reliability Estimates,” *British Journal of Contemporary Education*, vol. 2, no. 1, pp. 17–29, Feb. 2022, doi: 10.52589/BJCE-FY266HK9.

[25] S. Wahyuning, *Dasa-Dasar Statistik*. Semarang: Yayasan Prima Agus Teknik, 2021.
[26] R. Ananda and M. Fadhli, *Statistik pendidikan: teori dan praktik dalam pendidikan*. Medan: Widya Puspita, 2018. [Online]. Available: http://repository.uinsu.ac.id/id/eprint/3586
[27] H. Taherdoost, H. Business, S. Sdn, C. Group, and K. Lumpur, "Validity-and-Reliability-of-the-Research-Instrument-How-to-Test-the-Validation-of-a-Questionnaire-Survey-in-a-Research interesante revisar el concurrente.pdf," *International Journal of Academic Research in Management*, vol. 5, no. 3, pp. 28–36, 2016.
[28] F. Fatayah, I. F. Yuliana, and L. Mufidah, "Validity and Reliability Analysis in Supporting Mastery Learning STEM Model," *BP*, vol. 18, no. 1, pp. 49–60, Feb. 2022, doi: 10.36456/bp.vol18.no1.a5175.
[29] S. Ekolu O. and H. Quainoo, "Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods," *Global Journal of Engineering Education*, vol. 21, no. 1, pp. 24–29, 2019.
[30] A. Friatma and A. A, "Analysis of validity , reliability , discrimination , difficulty and distraction effectiveness in learning assessment Analysis of validity , reliability , discrimination , difficulty and distraction effectiveness in learning assessment," *Journal of Physics: Conference Series*, vol. 1387, no. 012063, 2019, doi: 10.1088/1742-6596/1387/1/012063.
[31] A. Kausar, S. Daimi, and T. Borulkar, "Assessing an assessment tool: Analysis of multiple choice questions on difficulty level and discrimination power, from an assessment in physiology," *Natl J Physiol Pharm Pharmacol*, no. 0, p. 1, 2021, doi: 10.5455/njppp.2022.12.103872202129112021.